International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979; Impact Factor: RJIF 5.72
Received: 23-09-2019; Accepted: 26-10-2019
www.allsubjectjournal.com
Volume 6; Issue 11; November 2019; Page No. 218-223

Machine learning approach for breast cancer diagnosis

Alqua Anjum
Research Scholar, Department CSE, SVIET, Banur, Chandigarh, India

Abstract
One of the most common cancers which is affecting the women worldwide is Breast Cancer. In world, today, Breast cancer has become one of the major problems causing the maximum cancer-causing deaths, and along with this, according to the global statistics, the majority of new cancer cases reported are breast cancer cases. The American Cancer society invests more in the research of Breast cancer than in the research of any other cancer. Early detection of this cancer is very difficult as it does not come with the clear figures, signs, and symptoms. However, only the early diagnosis of BC can help in improving the survival rates of patients as these will be provided with the timely and proper treatment. Further if its detected earlier and properly that whether the cancer is benign or malignant it can prevent the patients to undergo the tough stages of treatment like biopsies, which are very painful. It can be seen here that in order to treat the BC patients properly, much research and work to be done on the classification and prediction of BC cancer types is needed i.e; to find out whether the cancer is malignant or benign. Humans (even the medical professionals) can make mistakes in detecting the cancers, therefore here we require a computer based system which is properly trained and can detect the cancers perfectly with maximum accuracy. Because of its wide use and the most accurate and proper results of analysis, Machine learning has been chosen as a proper tool for classification and developing the predicting models to predict the cancer types. In this paper, we have used the different classification models of machine learning (kNN, decision tree, adaboost, Logistic Regression, Random Forest, bagging) and have provided their results along with the accuracy. We have further developed a hybrid data model- A Voting Ensemble Method. We have tried our best to combine the most compatible algorithms with each other so that it could provide us the more accurate results as possible. This voting ensemble method provided the accuracy of 90%. The maximum and most promising accuracy we came up with, was of Random Forest
99% with very good cross-validation score value. We draw our primary data from WISCONSIN BREAST CANCER DATABASE.

Keywords: FNA, breast cancer, machine leaning, feature selection

1. Introduction
The second major cause of the women's death in the world today is Breast Cancer and not only in women it has also put its deadly impact among men. The end of danger is not here but according to a research it has been found that it is the fastest growing cancer in our society. In 2017, 252710 new cases of breast cancer were reported among women and 2470 cases among men, among which 40610 women and 460 men died due to this cancer. Following table shows the estimated ne female breast cancer cases and deaths by age, in US, 2017 [1].

New female Breast Cancer Cases by Age recorded in US in the year of 2017,

	Insitu cases		Invasive cases		Deaths	
Age	number	%ge	number	%ge	number	%ge
< 40	1,610	3	11,160	4	990	2
40- 49	12,440	20	36,920	15	3,480	9
50- 59	17,680	28	58,620	23	7,590	19
60-69	10,550	28	68,070	27	9,420	23
70-79	10,370	16	47,860	19	8,220	20
80+	3760	6	30,080	12	10,910	27

Fig 1: Breast cancer case records in US

The proper diagnosis of the breast cancer has become need of an hour in the world today. Proper diagnosis requires the proper detection of cancer types in a person as early as possible. However it is not the easy task to find out whether a person is suffering from the malignant or benign cancer. Traditional methods include the biopsies, tests, regular check- ups and observations, etc which are very costly and time taking. These tests took a lot of time to provide the results, in which the chances of reduction of survival rates in patients may increase. Also the accuracy of results depends

upon the physician's expertise which is not guaranteed. The another drawback of the traditional method is that the patients with benign tumours need to undergo such painful and costly tests necessarily. The major role in predicting the cancers and hence in diagnosing the breast cancers can be played by the Information and Communication Technologies. The usage of Various MACHINE LEARNING TECHNIQUES used in different fields like healthcare , applied sciences rise rapidly due to their better performances, high rates of accuracy, reduced costs, reducing time. With the advancement in the technology it is very easy to store and obtain the data from the large number of sources and using machine learning techniques this data can be analyzed. Thus developing an intelligent system to classify the Breast Cancer tumours very efficiently. There are number of classification algorithms that can classify the tumours based on their features like SVM , logistic regression, decision tree, adaboost, etc. however the algorithm providing the accurate results are considered during research. This paper provides the performances of different classification algorithms included logistic regression, Random forest, adaboost, k-NN and various ensemble methods, which are among the most influential data mining algorithms. We have also showed the performance of a newly developed HYBRID data model, which shows comparatively a very good performance. The main motive of this research is to develop the different classification models and show their effective results of accuracy i.e to show how accurate they are and to find out the best one. We have also used the technique of Feature selection and showed its effect over the performance of respective algorithms.

2. Machine Learning Concept

Machine Learning is a new and most growing technology of data sciences. It is an approach where the machines are trained to mimic the decision making function of human brain. The functioning of Machine Learning involves the development of algorithms that allow the machines or computers to make the decisions based on the data obtained from different sources [2]. Machine Learning programming concept is different from the traditional programming. In case of traditional programs we define the rules and provide the required input and output is obtained while in case of Machine learning input and output both are provided in order to obtain certain rules for the functioning the machine.

a. Types of machine learning Algorithms

Unsupervised Machine Learning: It is a kind of learning where no labels are given that's here no teacher/supervisor is present here to train the machine. The prediction is based on the bases of previous experiences, patterns [3].

Supervised Machine Learning: This kind of learning is labelled based, that is here a teacher/supervisor is present to train the machine and hence enhance the results. In other words we can say here the input/output both are known for the number of instances and the machine is trained on the bases of this data. It is further categorized into two categories; Classification and Regression [3].

Our project is based on the supervised machine learning as we are using the data already provided with the class labels. This data along will be used to train the machine so that it could forecast the results for any input provided in future. The accuracy of the prediction will depend upon the algorithm used so the main aim is to find the algorithm or the combinations of algorithm that could predict the result with maximum accuracy rates.

3. Related Work

In 1994 a study was carried out under the name, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates" of which the breast cancer cytolic dataset was the part for first time [4]. A decision tree classifier was used here to train a model that could classify the cancers. And it gave the results with 95% accuracy [4]. Since then a number of studies were carried out that used the other classification algorithms to classify the cancers and the motive was to find out the best with least problems in handling the data.

Vikas Chaurasia *et al* [5], used a large dataset with 683 instances and applied the three machine learning algorithms including Naïve bayes, RBF network and j48. To remove the overfitting they used 10-fold cross validation. Among these three algorithms Naïve bayes came with the maximum accuracy of 97.37%, followed by RBF with 96.77% and then j48 with accuracy percentage of 93.41%.

D.Lavanya and Dr.K.Usha Rani [6] used the CART algorithm to classify the tumors. They used the concept of feature selection and showed the results both before and after selecting the specific features. In their work they showed the importance of feature selection and showed how the accuracy rates improve from case of no-feature selection to feature selection.

M. Tahmooresi [7] used a hybrid model in which they used SVM, ANN, K-nearest neighbor, and decision tree and successfully it came with the best result with the accuracy of 99.8%.

Y.Ireaneus Anna Rejani and Dr.S. Thamarai Selvi [8] used the SVM algorithm to classify the breast cancers. They used the various techniques like the Gaussian filtering technique to filter the images so that the relatable features could be taken before the real classification technique was used just to improve the accuracy of algorithm.

Roulan Xu and Qiongjia Xu [9] used five machine learning algorithms including Logistic Regression, Naive Bayes, Linear SVC, SVM SVM with linear kernel and Random Forest. They also used 3 feature selection techniques including PCA, RFE and HEAT MAP to select the most important features in a dataset. They concluded that the Random Forest and SVM came with the most accurate results with accuracy rates of 98% and 97% respectfully.

Peter Adebayo Idowu, Kehinde Oladipo Williams[10] used the 2 classification algorithms naïve bayes and j48 decision tree to classify the breast cancers. They concluded with the result that j48 has more accuracy than naïve bayes.

As we see a lot of work can be carried out over this topic. Now in our project we have first used the different classification algorithms including Random forest, k-NN, adaboost, logistic regression,bagging and compared their results to find out the most accurate algorithm. After that we write code for the voting ensemble method, where we tried to combine the different algorithms to develop a new model. We stopped adding further algorithms when we find out the maximum possible accuracy, that this model could give. We also used a feature selection method to select the most related features and showed its effect on the predicting accuracy. 5-fold cross validation technique is used to remove the over-fitting. The results, experiments and

observations of the experiment will be shown in the other sections of this paper.

4. Methodolgy

a. Environment and Languages

Due to the ease of learning involved in Python, it has today became one of the fastest growing programming languages in world. In comparison to the other languages, it's the available libraries that make python a successful language. There are some 72000 Python Package Index available and still growing. Pandas is a Python library used for everything in data analysis from importing excel sheets to processing datasets. Beside this it also provide libraries like Scilkit-Learn and PyBrain as machine learning libraries that provide tools for developing neural networks and data processing.

Other libraries include SciPy, NumPy (one of the oldest python tools successfully used, NumPy's functions are extended in Pandas and used for data analysis with advanced numeric analysis), csvkit, PyTables etc.

We have used Anaconda that provide different environments with different versions of Python and packages installed in them. It is very easy to create, export, list, remove and update different environments using conda.

b. Dataset Description

A medical diagnostic procedure namely "Fine Needle Aspiration", is a technique used to study the irregular lumps or masses on human body. Sample is taken from different patients which is separately examined under the microscope. After examining the images researchers obtain the different features for these cells and different feature values for different patients and hence a dataset is developed with at-least 32 features. It was developed examining 569 patients and is named as WISCONSIN dataset. Though during research a number of features were obtained but only the most effective 32 were chosen to be put into the database. These features describe the different characteristics of the cell nuclei studied from the images in 3-d space. 10-real values of the cell nuclei that are collected include 1) Radius (mean of the distances from centre to points on the perimeter) 2) Texture (standard deviation of grey-scale values) 3) perimeter 4) area

5) smoothness (local variation in the radius lengths) 6) compactness (perimeter^2 /area - 1.0) 7) concavity (severity of concave portions of the contour) 8) concave points (number of concave portions of the contour) 9) symmetry 10) fractal dimension ("coastline approximation" - 1) [11]

For the each of these real value features the Mean, Standard Error, and Worst or Largest (mean of largest values) resulting in 30 features and other two attributes include "id" and "label-diagnosis". There is no missing value for any attribute in this database. The Class distribution is as; nign and 212 malignant.

c. Data Pre-processing

Data preprocessing is the step were the data obtained is processed and the NAN values, missed values are removed in a dataset or are replaced with the new calculated values. In our project first step of data preprocessing involved label encoding so that to convert it in the format that could be understood by Python. After that we do splitting of data into the Train and Test data. We split our dataset in ration of 8:2, which means we have 80% of train data and 20% of test data.

5-fold cross validation technique is being used for scaling the data.

d. Classification Algorithms Used

Logistic Regression [12]: Using this Algorithm we can classify the dependent binary data. it can also classify multi nominal variables, interval level variables etc. Here a thresh hold is chosen against which the classification decision is made.

Ada Boost [13]: Ada Boost or ADAPTIVE BOOSTING is an ensemble technique of machine learning that combines the several weak classifiers to develop a stronger classifier in order to minimize the rate of error.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

K-Nearest Neighbor [12]: In this type of the classification model k points are selected and then all the data points are classified into the k-classes depending upon their distance from the point. The data point is place in that class from which its distance is least. This algorithm is used for both classification and regression techniques. Distance is measured using Euclidean Distance Formulla;

Random Forest [14]: Random Forest is a classification techniques where the number of Decision Tree classifiers are combined and merged together to get more accurate and correct results. It is flexible and easy to use method and produces results with good accuracy even without using any Scaling method to scale data.

Bagging [15]: Bagging is the short form for bootstrap algorithm. It is the ensemble algorithm that ensembles the various decision trees to provide its results. It helps in reduction of variance and also avoids over-fitting. It is basically designed to improve the accuracy of classification model.

Voting ensemble method [16]: it is the simple method of developing the predictive models by combining the different independent classifiers and predict the right results. In this algorithm the results depends on the number of votes for the particular result from different algorithms used.

5. Results and Observations

Classification algorithms discussed in previous sections, provide us with the different results. The results obtained during our research, are obtained in terms of accuracy. For some algorithms we also get the confusion matrices, to obtain the predictions of different classification algorithms in a simple and short method. We also got the other measures like F1 score, precision, recall of different algorithms, and in this paper we have compared the results of different algorithms by comparing these different measures obtained.

To show the impact of Feature selection method on the performance of different algorithms, we have first shown graphically the relationship between a respective feature and the target and then find out the 5-most effective features, and using these we find out the accuracy of algorithms. We have also used 5-fold cross validation technique to remove the over-fitting and have shown the cross validation scores.

Figure 2 and Figure 3, gives us the confusion matrix of KNN and LOGISTIC REGRESSION techniques along with their f1 score and accuracy. From the confusion matrix the

various other measures can be obtained like here we obtain the accuracy of two algorithms which is 95% and 96% respectively

```
   ...: print(accuracy_score(y_test,y_pred)*100)
[[76  4]
 [ 1 33]]
0.9295774647887325
95.6140350877193
```

Fig 2: Accuracy of KNN

```
   ...: print(accuracy_score(y_test,y_pred)*100)
[[76  4]
 [ 0 34]]
0.9444444444444444
96.49122807017544
```

Fig 3: Accuracy of LR

After applying the KNN and LR we tried to find out the effect of ensemble classification algorithms on our dataset. For this the first algorithm we used to predict the results is bagging method. In this method we used 20 estimators to make the predictions and provide with the best possible results.

The accuracy percentage provided by this algorithm is 88%. Though it is not more than that of KNN and LR, however it is expected to show better results with the larger datasets.

Now, another VOTING ENSEMBLE method is developed by combining the compatible algorithms and its effect over prediction of right results is calculated. Here we combined decision tree, svm, lr, and adaboost, found them more compatible and applied it. It provided us the accuracy of 90% which is more than that of Bagging.

Continuing our research, we now tried to find the correlation between the features and target, so that algorithm could find out the most effective features and will use only these in giving the predictions. This technique helps to show the effect of choosing the proper features in the dataset.

Figure 4, shows the histogram relationship between the some feature columns and the target column (malignant and benign) in dataset

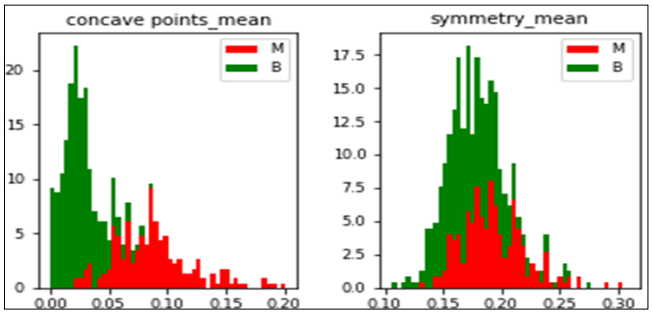

Fig 4(a): Feature Analysis

From this screen shot we can observe that larger the value of concave points –mean, shows the greater correlation with the malignant cancer, however symmetry mean does not show that much correlation i.e; it does not separate the two tumors properly.

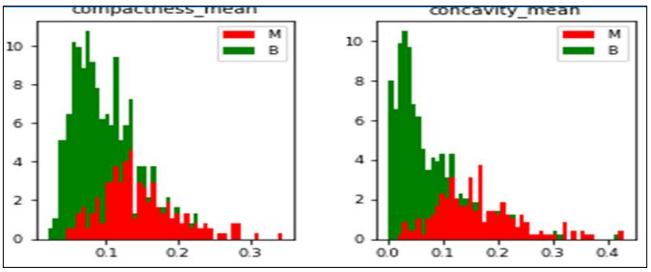

Fig 5(b): Feature Analysis

The above screenshots show the impact of increasing values of compactness and concavity on distinguishing the tumors. it is clearly seen that as the values are increased the impact on accurate separation is increased.

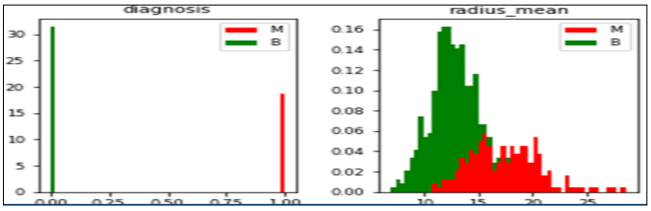

Fig 6(c): Feature Analysis

The above screenshot shows the impact of increasing values for radius mean on the detection of cancer –benign and malignant. We can see as the radius mean is increased the chances for a cancer to be malignant are increased.

The feature important series after finding the correlation results is found as is given in figure 5 as follows:

Table 1: Feature Importance Values

Features	Ranks
Perimeter-Mean	0.307
Area Mean	0.281
Radius Mean	0.274
Texture Mean	0.079
Smoothness Mean	0.057

After selection of the feature important matrix we run the different models over the dataset and obtained their accuracy along with their cross validation score. The different classification models used include LR, DT, and Random forest. We run the algorithms over 5-selected features as well as over single feature to observe the different behaviors.

While applying decision tree algorithm we observe over fitting, so its results can be ignored as these are not valid. It provided us 100% accurate results that could not happen.

Logistic regression provided the accuracy of 91%. However when applied over the single feature column the accuracy calculated was 88%.

At last we applied Random Forest with 100 estimators and it provided the accuracy of 94.4%, which is best among all. We also tried this algorithm over whole dataset and the results were great. It is found 99% accurate with perfect cross validation score.

Figure 6, shows the histogram representations of the obtained accuracies of all the classification algorithms used in our research. We can observe from histogram how decision tree is showing the 100% accuracy, which actually is not true result and is only due to over-fitting. However best accurate results are shown by random forest when applied over the whole dataset.
The figure is shown as below:

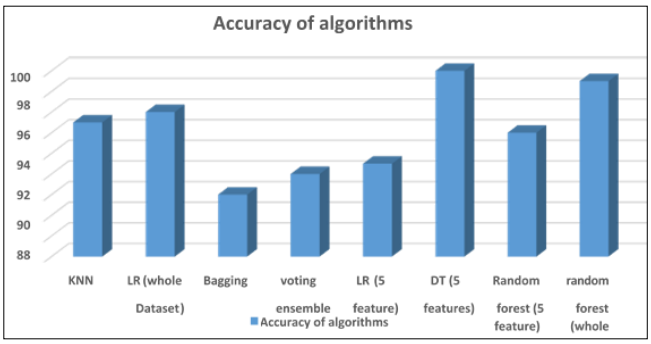

Fig 7: Comparison of various algorithms Used.

We have also calculated the other performance measures of certain algorithms that we used including KNN, LR, Voting ensemble method and Bagging. These measures include Precision, Recall and ultimately f1 score.
The ratio of correctly predicted positive values to the total predicted positive values is calculated under
Precision: As in the confusion matrix we obtain these things that is true positives and false positives, precision can be obtained as:

[TP/ (TP+FP)]= Precision
Recall or Sensitivity is the ratio of total number of true predictions for positive class to the total number of predictions for that very class. For example in our case, it is ratio of total number of true positives for malignant class to the total number of predictions made over this class. Can be obtained from confusion matrix as:

[TP/ (TP+FN)]= Recall
F1 score, it is the harmonic mean or the weighted average of Precision and Recall. It takes both TP and TN into account and calculates its value. It can be obtained as:

F1 score= 2* (Recall * Precision) / (Recall + Precision)
The Precision, Recall and F1 score of some algorithms used is shown and compared in following Figure 7:

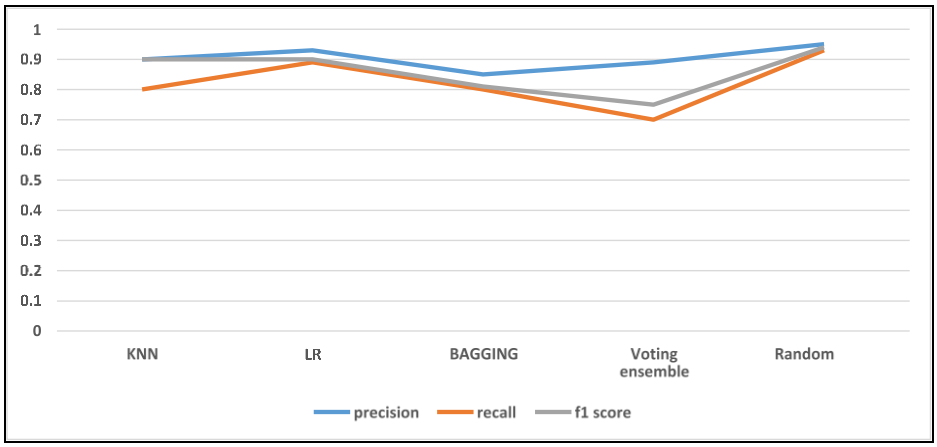

Fig 8: Precision recall and F1 score comparision

6. Conclusion and Future Work

In our research we have applied the different machine learning algorithms over Wisconsin dataset and in conclusion we find out that Random forest method of prediction works better than any other used. We tried to develop a new ensemble method by combining several algorithms, though its accuracy was good but not so satisfying as of random forest. However, we come up with the fact that how machine learning can play and important role in detection and diagnosis of Breast cancer.
In future the voting ensemble method we have used and be enhanced and hence a better model can be designed to predict the results. Also the dataset that we have used in our research is basic one , a new dataset with larger number of instances and feature columns can be used to develop and train the model. Further concept of deep learning can be introduced in the research and deep learning techniques can be utilized to develop prediction models.

7. References

1. Ruolan Xu, Qiongjia Xu. “Applying Different Machine Learning Models to Predict Breast Cancer Risk”J. Breckling, Ed. The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, 61.
2. Gayathri BM, Sumathi CP, Santhanam T. “Breast Cancer Diagnosis Using Machine learning algorithms–a Survey”, International Journal of Distributed and Parallel Systems (IJDPS), 2013, 4(3).
3. Sathya R, Annamma Abraham. “Comparison of Supervised and Unsupervised Learning Algorithms For Pattern Classification”, International Journal Of Advance Research in Artificial Intelligence. 2013; 2(2).
4. Ruolan Xu, Qiongjia Xu. “Applying Different Machine Learning Models to Predict Breast Cancer Risk”
5. V. Chaurasia, Surabh P. and BB Tiwari, “Prediction Of

Benign And Malignant Breast Cancer using data mining techniques", journal of Algorithms and Computer technology, 2013.
6. Dr. K Usha Rani, Lavanya D. "Ananlysis of Feature Selection with Classification: Breast Cancer Datasets", Indian Journal of Computer Science and engineering (IJCSE).
7. Tahmooresi M, Afshar Bashari, Nowshath Rad, Bamiah MA. "Early detection of Breast Cancer using Machine Learning Techniques", Journal of Telecommunication, Electronic and Computer Engineering, e-ISSN: 22898131, 10.
8. Ireaneus Y, Anna Rejani, Dr. S Thamarai Selvi. "Early detection of breast cancer using SVM"
9. Ruolan Xu, Qiongjia Xu. "Applying Different Machine Learning Models To Predict Breast Cancer Risk"
10. Adebayo Idowu P, Williams KO, Balogun JA, Oluwaranti AI. "Breast Cancer Risk Prediction Using Data Mining Classification Techniques", Transactions On Networks and Communications, Society of Science and Education, UK, 2015; 3(2).
11. Abien Fred, Agarap M. "On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset", ICMLSC, February 2–4, 2018, Phu Quoc Island, Viet Nam, 2018.
12. Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne, Xiaohui Liu. "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis", Designs, 2018, 2(13). doi:10.3390 /designs 2020013
13. Robert E Schapire, "Explaining adaboost", Princeton University, dept. of computer science, 35 Olden Street, Princeton, USA.
14. Marko Robnik-Sikonja, "Improving Random Forests", European Conference of Machine Learning, ECML, 2004, pp 359-370.
15. mailto:https://en.m.wikipedia.org/wiki/Bootstrap-aggregating
16. Jason Brownlee, "Ensemble Machine Learning algorithms in Python with scikit-learn" machine learning mastery, 2016.